LAMP-TR-116 October 2004
CFAR-TR-1001
CS-TR-4621
UMIACS-TR-2004-63

Exploring Interactive Relevance Feedback With a Two-Pass Study Design

Dina Demner-Fushman, Daqing He and Douglas W. Oard

Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742-3275
demner,daqingd,oard}@umiacs.umd.edu

Abstract

Interactive query refinement is widely believed to improve the effectiveness of ranked retrieval, but it can be difficult to leverage existing batch evaluation frameworks to quantify the relative benefits of alternative interaction designs. This paper uses the new two-pass interaction design of the Text Retrieval Conference's High Accuracy Retrieval from Documents (HARD) track to explore the design space for cluster-based interactive relevance feedback. Two sites contributed two techniques for cluster formation and three techniques for cluster labeling. The effectiveness of each technique was compared with lower bounds based on blind relevance feedback, and with upper bounds found with oracle-based techniques. The clustering techniques were found to yield potential benefits, but the automatically constructed cluster labels were found not to support sufficiently accurate cluster selection. Elicitation of a desired cluster descriptor was found to significantly improve the effectiveness of a subsequent retrieval pass. These results indicate that the affordable two-pass study design used in the HARD track can yield useful insights to guide future design decisions.

The support of this research by the Department of Defense under contract DARPA cooperative agreement N66001-00-2-8910 is gratefully acknowledged.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
OCT 2004		**00-10-2004 to 00-10-2004**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Exploring Interactive Relevance Feedback With a Two-Pass Study Design	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Language and Media Processing Laboratory,Institute for Advanced Computer Studies,University of Maryland,College Park,MD,20742-3275**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**20**	

1 Introduction

Information retrieval systems that seek to identify topically relevant documents based on a clear articulation of a searcher's information need face two key challenges in current search environment: (1) searchers often initially lack a clear understanding of their information needs, and (2) searchers, especially naive searchers, often lack the detailed understanding of the document collection to pose effective queries. As a result, searchers often initially pose short and poorly focused queries to interactive information retrieval systems [5, 7]. That is typically followed by some form of iterative refinement in order to converge on a query that yields a useful ranking of the retrieved documents. One widely studied approach to iterative refinement is interactive document-scale relevance feedback [3, 2]. This has been shown to be remarkably effective using automated experiments in which the relevance feedback step is simulated by designating highly ranked documents assumed to be relevant for feedback. Paradoxically, searchers have been observed to make little use of interactive document-based relevance feedback when provided with that option [7, 9].

Clearly, fully automatic evaluations fail to capture some important characteristics of real searcher-system interactions, such as the time and cognitive effort required to assess the relevance of several documents. Controlled user studies can be designed to characterize the effect of those factors, but traditional quantitative study designs require substantial investments. For example, the within-subjects design used to compare two systems in the Text Retrieval Conference required about three hours each from at least eight subjects [13]. Such an investment can be justified for summative evaluation at the end of a development cycle, but formative evaluation during development demands that more affordable ways of comparing a broad array of alternative interaction designs be found.

This past year, the Text Retrieval Conference (TREC) introduced a new track for evaluation of interactive query refinement, which is called High Accuracy Retrieval from Documents (HARD). In this paper, we are going to present our exploration of interactive relevance feedback under the framework of HARD, which provides a novel two-pass study

design with features different from both the automatic and the fully interactive evaluation frameworks.

In section 2, we first briefly introduce the HARD track, then we illustrate some of the insights about interactive relevance feedback that can be gained from this evaluation design in section 3, presenting results based on clustering highly ranked documents, automatic cluster labeling, interactive cluster selection, and automatic query refinement using the documents in selected clusters. This is complemented in section 5.4 with an exploration of ways to obtain additional interactive feedback, including situations where no clusters have been selected. We evaluate which of the following approaches should be used: requests for user-generated cluster labels, guided elicitation of salient terms, or document-scale feedback based on recently searched documents outside of the collection. We conclude the paper with the discussion of the strengths and weaknesses of the two-pass approach to formative evaluation of alternative interaction designs.

2 The TREC-2003 HARD Track

Relevance feedback has previously been explored using fully automatic experiments, observational studies, and controlled user studies. Automatic experiments leverage available relevance judgments, typically modeling interactive feedback as accurate designation of relevant documents that are highly ranked in a first retrieval pass [11]. A similar approach was also used to assess more selective feedback based on active learning in the Text Retrieval Conference (TREC)'s adaptive filtering task [10]. This type of fully automatic experiments offers affordable ways of quantifying the potential benefits of alternative techniques for nominating documents for feedback and for exploiting feedback that is received, but it reveals nothing about whether searchers will be able to accurately identify relevant documents in practice.

Remarkably, observational studies consistently indicate that most searchers rarely take advantage of interactive relevance feedback. This poses somewhat of a paradox; we know that relevance feedback is useful, but we also know that searchers don't take advantage of the capability. This might be explained in several ways. Perhaps searchers are not able to reliably recognize relevant documents. Perhaps they could accurately

recognize and designate them, but the time and effort needed to do so exceeds the perceived value. These factors (and others) probably explain a part of the problem. In this paper, we use the controlled user study design from the TREC-2003 "High Accuracy Retrieval from Documents" (HARD) track to explore the first of these factors, exploiting clustering and simple forms of multi-document summarization to minimize the demands placed on the searcher. This study permits exploration of the promising directions before embarking on traditional interactive study designs two of which were studied in the TREC interactive track. In the within-subjects user study design used in the TREC interactive track, each participant ("subject") tried two (or more) interactive systems, typically performing several searches with each [4]. Such a design can reveal statistically significant results with far fewer participants that the alternative "between subjects" design in which each searcher uses only one of the available systems. Both types of study design incur significant costs, however; reliably observing system effects for a simple two-way comparison in the presence of searcher, topic and presentation order interactions typically requires dozens of hours of closely observed searching by the participants.

The TREC-2003 HARD track introduced an affordable study design in which the focus is on richly modeling the information need and search context of a specific searcher. HARD exhibits the following key differences from the earlier TREC Interactive Track [1]:

(1) Interaction with the searcher is restricted to two episodes for each topic in the HARD track, the first when the query is posed and a second that can be based on initial search results. This two-pass design is sufficient to assess the benefit that results from a single relevance feedback opportunity. Interaction with the searcher is limited to three minutes per topic, which enables the use of either more topics or more system variants than would be possible with the interactive track's design. For the 2003 HARD track, interactions were further limited by requiring that they be accomplished using a static Web page and that they (generally) not require scrolling. Those technical limitations are not fundamental to the study design, however, and they may be relaxed in future years.

(2) Outcomes are determined by comparing measures of ranked retrieval effectiveness before and after feedback from the searcher becomes available. The use of a common

outcome measure can facilitate cross-site comparisons, and the use of automatically computed ranked retrieval measures allows alternative ways of exploiting searcher responses to be affordably evaluated. The fully interactive experiments, such as TREC interactive track , by contrast, usually adopt user-oriented outcome measures (e.g., aspect recall) that did not offer a similar potential for reuse.

(3) No attempt is made to control for interactions between searcher characteristics and presentation order. Each topic belongs to a single searcher; they pose the original query, and they interact once with each system to help refine that query. HARD track results therefore convolve learning and fatigue effects with the main (interaction) effect that we seek to measure. Key results from HARD track experiments should therefore be confirmed using more traditional study designs. Searcher effects can be controlled to some extent, however, because several searchers were involved (each working with different topics).

The document collection for the 2003 HARD track included 320,380 English news stories and 51,839 documents from two US Government sources, all from 1999. 48 topic descriptions (structured statements of an information need from which queries can be automatically defined in several standard ways) were also provided. Subsequent interaction with the searcher was accomplished using automatically generated "clarification forms." Clarification forms typically asked the searchers to make some selection (e.g., indicating whether they felt a summarized set of documents would meet their needs) and/or to enter some text. Binary topical relevance judgments (relevant or not) were later created for the top 75 documents from each participating system using a pooled assessment methodology. An average of 160 relevant documents were found for each topic (ranging from 6 to 714). Traditional interactive relevance feedback techniques require the searchers to indicate which documents they have found to be useful (and sometimes, not useful). Real searchers generally assess documents based on a broader range of criteria than topical relevance, so a more selective set of utility judgments was therefore also provided to indicate whether topically relevant documents were appropriate for the intended purpose and pre-existing degree of familiarity, and were drawn from the preferred genre. The

relevance and utility judgments were usually made by the same searchers that created the topic description and interacted with the clarification forms. Mean Average Precision (MAP) is reported below for both sets of judgments; the majority of our analysis focuses on topical relevance.

3 Interactive Relevance Feedback

Interactive relevance feedback is different from the automatic relevance feedback because the searcher who issued the query is involved in the feedback process. The advantage of human involvement is that the human's world knowledge and ability to reason and recognize patterns can be leveraged to compensate for the system's weakness in those areas. However, human's weaknesses, such as inconsistency and short attention period, should also be considered.

One of the disadvantages of the interactive document-based relevance feedback is that viewing entire documents can require substantial time and screen space. Studies have shown that Web searching involves shorter sessions and fewer iterations than is typically seen in more traditional search tasks (e.g., searching a library catalog) [6]. A commonly used approach to reduce the load of evaluating whole documents in short sessions, is to allow searchers to designate (hopefully) relevant documents based on brief summaries (e.g., titles).

In this study, we have taken this one step further. We want to explore the approach of displaying brief multi-document summaries for clusters of topically similar documents. By limiting the number of cluster summaries that are displayed, we hope to reduce the cognitive effort associated with interactive relevance feedback and simultaneously minimize the time and effort associated with selection. Our approach involves four steps 1) automatic clustering, 2) automatic cluster labeling, 3) interactive cluster selection, and 4) automatic query expansion.

Furthermore, with searcher's involvement, we could ask not only for the relevance judgments on the documents selected by the retrieval system, but also for extra information that the searcher could provide about the search topic. The elicited information ranging from a set of relevant terms to additional relevant documents was used in our

experiments as an alternative to the cluster-based query expansion.

4 Experiment Design

Our experiments were designed within the HARD framework. The first episode for each of the 48 topics was an ad hoc retrieval. Queries were provided in the form of TREC-style topic descriptions, and we only used "title queries" (all of the words from the TITLE field of the topic description) because the title field in TREC topic descriptions is intended to be representative of the relatively brief queries that searchers often pose to Web search engines. Our title queries had an average length of three words. Examples ranged from well articulated needs, e.g. "NATO/UN Tension over Balkans Crisis" to non-specific requests such as "Child's play." Topic descriptions for the HARD track include several additional fields designed to support searches based on several aspects of utility beyond simple topical relevance, but with the exception noted in section 5.4 we did not use the contents of those fields for the experiments reported in this paper.

As stated, our goal is to explore interactive relevance feedback within the two-pass design. We concentrated on studying the approaches for cluster-based relevance feedback, and the approaches for eliciting extra information. Through the experiments, we want to answer the following questions regarding cluster-based relevance feedback:

- How to perform clustering for interactive relevance feedback? How many documents to use? Can current clustering algorithms generate useful clusters?
- How to label clusters? Can searchers select relevant clusters based on the automatically generated labels?
- How to present the clusters and their labels? What presentation order of the clusters permits obtaining accurate answers and avoiding learning and fatigue effects?
- How to utilize the answers in automatic query expansion?

These four groups of questions lead to natural organization of our experiments on cluster-based relevance feedback into the following four sections:

For the experiments exploring automatic clustering, we tested two approaches. In the approach, labeled "experiment 1" in this paper, University of Massachusetts Amherst performed an initial title-query search using Lemur[1], then formed clusters from the top 200 documents using group average clustering, and selected top 15 clusters according to the size of the clusters. In another approach, we generated the initial title-query search results using InQuery text retrieval system (version 3.1p1) from the University of Massachusetts, and then used Ward's method [8] to cluster the top 10 documents. This typically resulted in 4-5 clusters.

As for the approaches to labeling clusters, in experiment 1, University of Massachusetts displayed the title of the document that was most similar to the cluster centroid, plus the top 10 terms from the cluster. We explored two other approaches. The first one was to display just the title of the document that was most similar to the query. The second one was to display a multi-document summary that was automatically generated by GOSP, an extractive summarizer trained on a corpus of daily news stories and headlines of weekly summaries [14]. These labels contained an average of 18, 13, and 17 words, respectively. Our first labeling approach, combined with our approach to clustering mentioned above, is labeled "experiment 2" in this paper. Our second approach is labeled "experiment 3".

Interactive cluster selection was performed by the HARD track relevance assessors at Linguistic Data Consortium (LDC)[2] using topic-specific clarification forms. Three sets of clarification forms were completed, each corresponding to the above cluster labeling approaches respectively. Because this part was performed at LDC, which is out of our control, we will not discuss this part further, and assume that all the answers were obtained under equal experimental conditions and have equal bearing on the results of our experiment.

Query expansion was performed automatically based on the answers generated in interactive cluster selection. Three forms of answers were obtained via searchers' selection of the clusters: terms, cluster labels, or documents. If the answers were terms or

[1]http://www-2.cs.cmu.edu/~lemur/
[2]http://www.ldc.upenn.edu/

labels, we added all the terms, or all terms from the selected labels to the original query, thus constructing the expanded query for the second search episode. If the answers were documents, we merged all terms from all documents in every selected cluster(after stemming and removing stopwords), ranked those terms in decreasing order according to their weights which were calculated by $tf_i * idf_i$ where tf_i is the number of occurrences of the i^{th} term in the merged documents, $idf_i = \log_2(\frac{N}{df_i})$, N is the number of documents in the collection, and df_i is the number of documents in the collection that contain the i^{th} term. We selected top 10 terms for query expansion.

Our studies used automatic blind relevance feedback (BRF) as a baseline. This comparison would tell us whether or not the cluster-based interactive relevance feedback is more useful than automatic query expansion, which has been demonstrated to be a useful technique for improving retrieval results. We used the top 10 documents from the initial search to expand the query.

To answer the questions about how useful the clustering algorithms are and how useful the cluster labeling techniques are, we employed several runs based on post-hoc analysis of the relevance judgment data. These runs, referred to as *oracle runs*, helped us establish various **upper bounds** for studying clustering, cluster labeling, and query expansion methods. They are, to some degree, extrinsic evaluations of these methods using the retrieval effectiveness (e.g. MAP) as the measure. We established four such upper bounds, which are:

Aggressive Selection. Here, we assumed that 1) the labels of the clusters could accurately reflect the content of the clusters, and 2) the searcher took a recall-oriented strategy when selecting the cluster. So, every cluster that contained at least one relevant document would be selected in this run. Obviously, this recall-oriented strategy is typically suboptimal, but since it can be achieved in reality under the above conditions, and reflects the selection strategy we suggested to the searchers, we kept it as the weak oracle run in the study of cluster labeling approaches.

Optimal Selection. Here, again, we assumed that the labels of the clusters could accurately reflect the content of the clusters. But, the assumed strategy for the searcher

was to select the optimal combination of clusters that would yield the best retrieval effectiveness (measured by MAP). This is the best feedback that we could hope for from a searcher, so the difference between Optimal Selection oracle run and the real Interactive Selection from a searcher is a measure of the degree to which the searcher was able to exploit the selection cues that we provided in the cluster labels. Due to the relatively large number of clusters, Optimal Selection proved to be computationally intractable for experiment 1. We tried every possible combination of clusters for experiments 2 and 3, and selected the set that yielded the best MAP with topical relevance judgments.

Optimal Clustering. Just as oracle runs with an ideal cluster label assumption can help us to study the effect of cluster labeling, this oracle run is a run with an ideal clustering assumption, which helps us to study the effect of clustering approaches. This run assumes that 1) we have a clustering algorithm that can accurately identify every relevant document above the clustering cutoff (e.g. 10 and 200 documents), and form two clusters, one with all the relevant documents, and the other with non-relevant; and 2) the searcher can identify the cluster with all the relevant document. The MAP difference between Optimal Clustering run and Optimal Selection is a measure of the usefulness of clusters generated by an automatic clustering technique.

Perfect Feedback. This run actually reflects the performance of a retrieval system. It assumes that every relevant document within the collection can be identified, and can be clustered into one cluster. Then, the searcher can select this cluster. The difference between Perfect Feedback and Optimal Clustering is a measure of the degree to which the clustering cutoff adversely affected the potential usefulness of cluster-based relevance feedback.

Cluster-based interactive approach poses questions complementary to the ones directly concerning clusters, regarding the need to ask for additional information, when e.g. no clusters were selected as relevant. The two-pass design gave us the opportunity to ask the searchers for extra information, which cannot be studied in automatic rele-

vance feedback. However, because the design did not permit the follow-up discussion of the elicited information with the searcher, it is an interesting research question to study what type of extra information to elicit from searchers under these circumstances, and how to use such information in feedback process.

In our experiments, we explored the options of asking for the following three types of extra information:

1. a short summary of the relevant cluster, about the same length as the cluster labels, when the searchers felt that no provided cluster was relevant based on their judgments.

2. a set of terms that the searchers thought were relevant to the search topics. The terms we specifically requested were person names, organization names, locations and any other terms that the searchers would like to provide.

3. any relevant document that the searchers saw previously.

When the elicited information was either short summaries or key terms, we treated them the same way as we did the cluster labels in query expansion. Whereas, if the elicited information was a document or an URL of a documents, we first removed the cases where only broad root directories (e.g. www.khoj.com) were available, then retrieved the documents, and striped HTML markup in the documents. Then we extracted relevant terms from the cleaned documents using the term extraction method described above.

5 Results Analysis and Discussion

5.1 Automatic Clustering

Our intrinsic analysis of clustering algorithms indicated that both approaches used in the RFBR' experiments and in our experiments generated reasonable clusters. As shown in figure 1, our approach, which used top 10 documents for clustering and Ward's algorithm, generated much higher percentage of clusters with high proportion of relevant documents than the RFBR's approach using top 200 documents. The quality of our clusters can

also be seen in the difference between our Optimal Selection run (see row "2+3 Optimal Select" in Table 1) and the corresponding Optimal Clustering run. The Optimal Selection run, which is the best performance given our clustering approach, has only 0.6% relative decrease in topical MAP compared to the performance of a perfect clustering algorithm which will cluster all retrieved relevant documents together into one cluster.

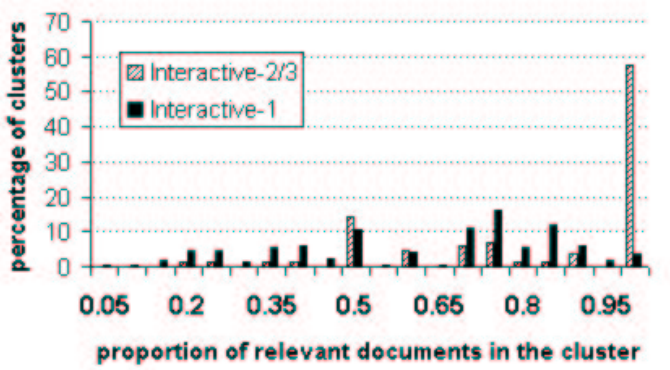

Figure 1: Density of relevant documents in clusters that contained at least one relevant document.

Viewing the clustering approaches extrinsically, i.e., looking at the MAP achieved by these runs, the 10-document clustering cutoff, however, turned out to be too restrictive. These runs resulted in a statistically significant($p < .01$) 13% relative decrease in topical MAP for Optimal Clustering when compared to the 200-documents threshold used in experiment 1.

The larger number of less pure clusters in experiment 1 clearly posed somewhat greater demands on the searchers, as is evident in the statistically significant($p < .01$) 7% decrease in topical MAP between Aggressive Selection and the best interactive result.

We also observed that aggressive selection strategy, which can guarantee to include all relevant documents in the clusters, but is aslo directly affected by the cluster quality, actually achieved better MAP results in experiment 1 than that in experiment 2 and 3. It seems that cluster quality in both experiments 1 and 2/3, to some degree, are good enough for searchers to perform reasonable cluster selection. However, we do know that there is still room for improvement in the clustering algorithm. The larger number of low quality clusters in experiment 1 clearly posed somewhat greater demands in the interactive relevance feedback process, as demonstrated by the statistically significant($p < .01$)

7% decrease in topical MAP between Aggressive Selection and the best interactive result of experiment 1. There is no statistically significant decrease in topic MAP between the correspoinding runs in expeirments 2 and 3.

Although both Aggressive Selection runs generated better results than blind relevance feedback run (BRF), only the experiment 1 achieves a 16% statistically significant improvement ($p < .01$) in topical MAP.

It seems that, to some degree, it is more important to identify more relevant information than to worry about bringing in the noisy information along with the relevant information. The relevance feedback mechanism is relatively robust to noisy information. From these results, we conclude that both clustering techniques yielded useful clusters, but that the 10-document cutoff that we used with Ward's method offered little potential for improvement over the BRF baseline.

Condition			Topical	Utility
Exp	Title Queries		0.3110	0.2617
All	BRF	Docs	0.3495	0.3033
1	Interactive	Label	0.3480	0.3274
1	Interactive	Docs	**0.3669**	0.3259
1	Interactive	Label+Docs	0.3629	**0.3450**
1	Aggressive Select	Docs	0.4054	0.3572
1	Optimal Cluster	Docs	0.4196	0.3865
2	Interactive	Label	0.3318	0.2988
2	Interactive	Docs	**0.3590**	**0.3205**
2	Interactive	Label+Docs	0.3454	0.3112
3	Interactive	Label	0.3132	0.2883
3	Interactive	Docs	**0.3478**	**0.3073**
3	Interactive	Label+Docs	0.3317	0.3035
2+3	Aggressive Select	Docs	0.3632	0.3244
2+3	Optimal Select	Docs	0.3644	0.3322
2+3	Optimal Cluster	Docs	0.3666	0.3350
All	Perfect Feedback	Docs	0.4492	0.4220

Table 1: The MAP of runs for experiments 1, 2, and 3, and the four oracle runs.

5.2 Cluster Labeling

How to label clusters is not an issue in the automatic relevance feedback studies, but it is in the study of interactive relevance feedback, especially in the framework of two-pass interaction design, where there is no chance to perform the follow up confirmation.

No interactive experiment achieved a statistically significant improvement over the BRF baseline. However, our oracle runs show that aggressive selection of clusters generated in experiment 1 could significantly improve the search result over BRF baseline. This discrepancy leads us to believe that the suboptimal cluster selection by the searchers affected the performance. Indeed, we observed how well searchers did at selecting clusters as shown in Table 2. In general, searchers selected fewer clusters than would have been useful. The possible reasons for the suboptimal performance could be 1) insufficiently insightful cluster labels, 2) an inappropriate selection strategy adopted by the searchers, or 3) simply mistakes happened in the selection process (e.g., too tired or not enough time to select all intended clusters). We could not interview the searchers, so we do not know which reason triggered the observed effects.

There were consistent improvements on both topical and utility MAP in all interactive runs of experiment 2 over that of experiment 3, which only differ in the cluster labeling approaches. It seems that labeling the clusters with the human generated sentences (e.g., the title of the centroid document) might be preferred in the process, even though multi-document summaries used in experiment 3 draw information from all documents in the cluster, whereas the title used in experiment 2 was from only one document.

Since the difference between the results of these experiments was not statistically significant, we cannot make a strong claim here. Another interesting observation in Table 2 is that the searchers selected clusters as relevant very conservatively, and thus did well in recognizing cases where selecting fewer clusters was appropriate. For example, in experiment 2, the searchers selected no clusters in 10 of the 13 cases where no relevant documents existed above our 10-document clustering cutoff (and in 9 other cases). The results in Table 1 are based on the use of title queries alone in those cases.

Interactive					Optimal Selection			
Exp	TP	FP	TN	FN	TP	FP	TN	FN
Clusters								
1	161	67	340	182				
2	49	3	85	36				
3	43	6	82	42				
Documents in Clusters								
2	143	39	216	102	136	31	224	109
3	123	27	228	122	136	31	224	109

Table 2: Cluster selection details. TP/FP: True/False Positive, TN/FN: True/False Negative.

5.3 Query Expansion

Two possible sources of information can be used for query expansion in interactive query refinement. The first one is the terms extracted from the documents in the selected clusters. This approach has been demonstrated to be useful in automatic relevance feedback, but its weakness is in poor performance when there is not much relevant information to extract from. The second source of information is the label of the cluster, especially when the labels are the titles of the centroid documents of clusters, which is different from the extracted salient terms. There are good reasons to use this second information source in interactive query refinement, since it is the label that the searcher's selection is based on. However, as these titles usually come from one document inside the clusters, they could be partial representation to what inside the cluster.

We did observe interesting differences in individual cases. For example, using labels alone typically was best when clusters are noisy or when the labels did not reflect the content of the cluster accurately. However, there was no statistically significant difference among the approaches of using labels, documents, or a combination of the two. The 3% difference in topical MAP between the best interactive results for experiments 2 and 3 was also not statistically significant, so we were unable to detect an effect from different cluster summaries.

5.4 Elicitation of Additional Information

The results of elicitation for extra information are shown in Table 3. Looking at both topical relevance and utility relevance, eliciting terms performed the best among all approaches, followed by eliciting label-like summaries. Eliciting documents, although seemingly obtained much more information from the searchers, performed the worst. The difference between eliciting terms and eliciting documents was statistically significant, so was the difference between eliciting terms and not eliciting any information ($p < .01$). No significant differences were found between eliciting short summaries and terms, or between short summaries and documents.

From this, we conclude that if we wish to minimize the demands to the searcher, our present feedback technique will benefit most if we ask the searcher for additional query terms when no useful clusters are found. Reliably detecting this situation automatically would be challenging, but the cluster labels appear to be sufficiently informative to allow the searcher to make that determination. It is a complex process to clean up the documents, especially those obtained from the Web. Therefore, the slight improvement gained using the elicited documents over not eliciting may not compensate for the effort.

Ask for	Topical	Utility
Nothing	0.2852	0.2209
Documents	0.2786	0.2653
Labels	0.2957	0.2733
Terms	0.3750	0.3469

Table 3: Effect of alternate elicitation strategies on experiment 2, for which no clusters were selected.

5.5 Combination of Feedback and Elicitation

Cluster-based interactive relevance feedback would essentially have no feedback information to use for query expansion when no cluster was selected by the searchers. In this case, elicited extra information would fill the blank of relevant information. We, therefore, explored one combination model that applies information obtained from elic-

itation when we know that no cluster-based interactive relevance feedback information was available to use.

results and diagram

6 Conclusions

In this paper, we discussed the study of interactive relevance feedback in the context of two-pass interaction design. We concentrated on examining cluster-based approach and the approaches of eliciting extra terms/documents. By augmenting the HARD framework with a set of oracle-based upper bounds, we were able to compare the effect of alternative clustering cutoffs, alternative techniques for clustering and for generation of cluster labels, and alternative elicitation strategies. Our results indicate that the two-pass interaction framework in the TREC HARD track provides a useful basis for examining some aspects of interactive retrieval and for sharing results across sites. And the insights we obtained are:

- Cluster-based interactive relevance feedback can generate reasonable improvement. Current clustering algorithms can produce relatively good clusters among top N documents. To realize the potential of the method and improve the retrieval effectiveness, it is necessary to select more than 10 documents for clustering.

- Labeling clusters using a set of representative terms was not as informative as labeling with summaries generated by machines, which was in turn not as good as labeling with the human generated sentence extracted from the centroid document of a cluster. We observed an aversive impact of the labeling, no mater which method was used.

- Eliciting extra information is an effective method for improving search results. Among the possible alternatives, asking for extra terms generated better retrieval results than asking for extra sentences or documents. These extra terms, if properly combined with the selection of clustered documents, could generate significant improvement over blind relevance feedback. These extra terms included person names, locations, and any terms that were important to the users.

- Eliciting extra documents introduced lots of extra processing work but little gain in retrieval effectiveness.

The two-pass interaction design does not limit us to study only the cluster-based relevance feedback. Actually, the interaction design for the HARD track can be viewed as a simplified model of the information need negotiation process described by Taylor [12]. Our study in this paper only concentrated on one of four key factors[3]. We reported some initial results about the other three factors in our HARD report (RFBR) Therefore, one of our future work directions is to study the other three factors in this framework, so that we can build a more comprehensive model about the interactions between a searcher and a retrieval system.

The labeling methods used in our experiments do not always represent the underlying clusters clear enough. In the future, we plan to explore possibilities of better labeling, guided by our observations that searches tend to make better selections when presented with the coherent concise cluster representations. The second area that we want to explore is that the difficulty of the query might be the key factor in selection of the interaction mode with the searcher during the second pass. We plan to explore the adaptive interaction in the context of query difficulty determination.

The last, but not least, direction of our future work is to find ways to compensate the limitation of using utility relevance in the design. We would see value in establishing some control over presentation order effects in future instances of the HARD track, and in establishing interchange formats that would encourage easy reuse of data collected at multiple sites. But it has been said that a journey of a thousand li begins with a single step; viewed in that light, the TREC-2003 HARD track was certainly a step in the right direction.

[3]The four key factors that human intermediaries (e.g., reference librarians) can productively discuss with searchers are: (1) the topic in which they are interested, (2) the characteristics of the searcher (e.g., their present knowledge of the topic), (3) their expectations regarding where the information they seek might be found, and (4) their preferences regarding the nature of their search results (e.g., do they seek an introductory treatment of the topic, or specialized technical details).

References

[1] J. Allan. Hard track overview in trec 2003 high accuracy retrieval from documents. In *The Twelfth Text Retrieval Conference*, 2003.

[2] E. N. Efthimiadis. Interactive query expansion: A user-based evaluation in a relevance feedback environment. . *JASIS*, 51:989–1003, 11 2000.

[3] D. Harman. Relevance feedback and other query modification techniques. In W. B. Frakes and R. Baeza-Yates, editors, *Information Retrieval: Data Structures and Algorithms*, pages 241–263. Prentice Hall, 1992.

[4] W. Hersh and P. Over. Interactivity at the text retrieval conference (trec). *Inf. Process. Manage.*, 37(3):365–367, 2001.

[5] P. Ingwersen. Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 101–110. Springer-Verlag New York, Inc., 1994.

[6] B. J. Jansen and U. W. Pooch. Web user studies: A review and framework for future work. *Journal of the American Society of Information Science and Technology*, 52(3):235–246, 2000.

[7] M. B. J. Jansen, A. Spink, J. Bateman, and T. Saracevic. Real Life Information Retrieval: A Study of User Queries on the Web. *SIGIR FORUM*, 32(1):5–17, 1998.

[8] A. Leuski and J. Allan. Lighthouse: showing the way to relevant information. In S. F. Roth and D. A. Keim, editors, *Proceedings of IEEE Symposium on Information Visualization (InfoVis'00)*, pages 125 – 130. IEEE Computer Society, October 2000.

[9] M. Magennis and C. J. van Rijsbergen. The potential and actual effectiveness of interactive query expansion. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 324–332, Philadelphia, 1997.

[10] S. Robertson. Threshold setting and performance optimization in adaptive filtering. *Inf. Retr.*, 5(2-3):239–256, 2002.

[11] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4):288–297, 1990.

[12] R. S. Taylor. Question-negotiation and information seeking in libraries. *College & Research Libraries*, 29:178–94, 1968.

[13] A. Turpin and W. Hersh. Why batch and user evaluations do not give the same results. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 225–231, Aug. 2001.

[14] L. Zhou and E. Hovy. A Web-Trained Extraction Summarization System. In *Proceedings of the HLT-NAACL conference*, May 2003.